Census 2000 Alternative
Questionnaire Experiment
December 12, 2002

Questionnaire Effects on Reporting of Race and Hispanic Origin: Results of a Replication of the 1990 Mail Short Form in Census 2000

FINAL REPORT

This evaluation reports the results of research and analysis undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation (TXE) Program, designed to assess Census 2000 and to inform 2010 Census planning. Findings from the Census 2000 TXE Program reports are integrated into topic reports that provide context and background for broader interpretation of results.

Elizabeth Martin

Directorate for
Methodology and
Standards

USCENSUSBUREAU
Helping You Make Informed Decisions

CONTENTS

LIST OF TABLES

EXECUTIVE SUMMARY

In 1997, the Office of Management and Budget introduced significant changes in methods for collecting and reporting race data in government surveys and censuses, including allowing respondents to report one or more races, and reversing the sequence of the race and Hispanic origin items. Other changes in format, categories, and wording were also introduced in Census 2000. In order to evaluate the net effects of all the changes, 1990 questions on race and Hispanic origin were replicated in a national experiment conducted during Census 2000.

During Census 2000, the Alternative Questionnaire Experiment 2000 mailed 1990-style short forms to an experimental sample of 10,500 households. The 1990-style form preserves 1990 question wording, categories, order, and format, but incorporates some recognizable elements of the 2000 design. A control panel of about 25,000 households received Census 2000 questionnaires. Mail return rates were very similar for both panels (72-73 percent). All experimental data were keyed and processed separately from the production census. For this report, data for both forms were edited by applying a simplified version of the pre-edits used in Census 2000 production. Missing data were not imputed or allocated, as they would be in fully edited census data. Results reported here may differ for fully edited and imputed data. Results of the experiment are generalizable only to the Census 2000 mailout-mailback universe. Excluded are mail nonrespondents enumerated in nonresponse followup, and segments of the population enumerated in other operations (such as American Indians on reservations and Alaska Natives).

Comparisons of results from the two panels show that Census 2000 questionnaire changes substantially improved the completeness of race and Hispanic origin reporting in mail questionnaires. Item nonresponse (i.e., blank or uncodable responses) for Hispanic origin was 3.33 percent in Census 2000-style questionnaires, compared to 14.46 percent in 1990-style questionnaires. Item nonresponse for race was 3.27 percent in Census 2000-style questionnaires, compared to 5.95 percent in 1990-style questionnaires. (For Hispanics, the reduction in race item nonresponse was very large, from 30.53 percent to 20.79 percent in 2000-style questionnaires.)

The Census 2000 questionnaire design also affected race reporting. Not surprisingly, reports of two or more races more than doubled (.82 percent to 2.03 percent) in response to the “mark one or more” instruction. There were more reports of Native Hawaiian and Other Pacific Islander race, and fewer reports of Some other race. Contrary to what might have been expected, there is little evidence that allowing respondents to report more than one race reduced single race reporting in the 5 major race categories (White, Black, American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islander). The exception is a reduction in reporting of White by non-Hispanics.

The effects of questionnaire changes on Hispanic race reporting were substantial. Reporting as White was higher by about 10 percentage points (48.98 percent, compared to 39.88 percent), and

reporting as Some other race was lower by the same amount (39.03 percent to 51.47 percent), in Census 2000-style forms. This result is consistent with prior research and probably reflects the effect of the new "one or more" option and the reversed item sequence. The results confirm the vulnerability of Hispanics' race reporting to question order and context effects.

Despite the reversed sequence of Hispanic origin and race and question wording differences, the same percentage (slightly over 11.1 percent) reported as Hispanic in both forms. This result implies that any changes from 1990 to 2000 in the fraction of the population identifying as Hispanic are not due to changes in design of the mail questionnaire. However, there were questionnaire effects on reporting of detailed Hispanic origin. The 2000-style questionnaires elicited fewer reports of specific Hispanic groups, and more reports of general Hispanic identity (e.g., Hispanic, Latino, Spanish) than the 1990-style questionnaires.

Comparisons of 1990 and 2000 census data must take into account the confounding effects of questionnaire changes on race reporting. For example, the changes in the design of the mail questionnaire would result in an increase from 1990 to 2000 in Hispanics' reporting of White race, and a decline in reporting of specific Hispanic groups, *even in the absence of any true changes in the racial or ethnic composition or identifications of the population*. These questionnaire effects may mask true population changes, or may masquerade as change when none has occurred.

Recommendations:

- **Conduct additional research into the reliability and causes of differential form effects on race reporting by Hispanics and non-Hispanics.**
- **Conduct additional research to develop more robust race measurement methods that are less vulnerable to methodological effects, especially for Hispanics.**
- **Conduct experimental research to evaluate the effects of other methodological influences on race reporting, including mode of interviewing and interviewer effects.**
- **Conduct additional research on the effects of examples on race and Hispanic reporting.**
- **Conduct research on the effects of changes in coding, pre-editing, editing, and imputation procedures on the comparability of race and Hispanic data**.
- **In future censuses, conduct replication studies embedded in the census to evaluate and calibrate the effects of questionnaire design changes (or other important changes in methods) on short form and long form data.**

1. BACKGROUND

In 1997, the Office of Management and Budget (OMB) introduced significant changes in methods for collecting and reporting race data in government surveys and censuses, including allowing respondents to report one or more races. In order to evaluate the effects of the OMB changes and other changes introduced in Census 2000, 1990 questions on race and Hispanic origin were replicated in a national experiment conducted during Census 2000. Data from 1990-style and Census 2000-style mail questionnaires are compared to address two questions.

- Does mail response data quality (as measured by item nonresponse) differ between questionnaire versions for race and Hispanic origin items?

- What are the effects of questionnaire differences on race reporting? Do race and Hispanic origin distributions for mail returns differ between 1990 and 2000 versions of the questionnaire?

The most significant change in Census 2000 was to *allow reporting of one or more race*s. The change culminated several years of research and consultations and a large national field test that evaluated alternative question formats (Census Bureau, 1997; Gerber, de la Puente, and Levin, 1998). Based on the research, the instruction was modified. The Census 2000 question is, "What is this person's race? Mark [✗] one or more races to indicate what this person considers himself/herself to be." (The 1990 census had asked, "Race. Fill ONE circle for the race that the person considers himself/herself to be.") The anticipated effect of the change is increased reporting of two or more races, and (possibly) reduced reporting in single race categories.

In 1990, race was followed (two items later) by Hispanic origin. A second major change in Census 2000 was to *reverse the sequence of race and Hispanic origin questions*. (This change is also required by the new OMB guidelines.) Research showed that when race came first, some Hispanic respondents looked for, but did not find, a category to identify themselves in the race question, and so reported "Other race" and wrote in a Hispanic group (see, e.g., Kissam, Herrera, and Nakamoto, 1993). The sequence also affected nonresponse to the Hispanic origin item, which was skipped by many non-Hispanic respondents who apparently thought it was redundant or did not apply to them. (In 1990, most people who skipped Hispanic origin were non-Hispanics; McKenney et al., 1993.) In order to address these problems, the Census Bureau in 1987 began experimenting with reversing the item sequence (Martin, DeMaio, and Campanelli, 1990). Asking Hispanic origin first would reduce the apparent redundancy, and allowing Hispanic respondents to first report their Hispanic identity would reduce the likelihood they would report it again in the race item. Several national field tests confirmed that reversing the order and adding an instruction to answer both questions reduced Hispanic item nonresponse by half, on average (Bates et al., 1995; see also Census Bureau, 1996; 1997). The reversed sequence also reduced Hispanics' reporting of Some other race. In Census 2000, Hispanic origin preceded race and an instruction to "Please answer both questions..." was added.

A third major set of changes involved the *format of the questionnaire*. Extensive developmental work and cognitive testing were conducted to improve the user-friendliness of the mail questionnaire. The matrix format used in 1990 was replaced with a columnar, individual space format, the separate roster of household members was eliminated, and white space and contrasting color background were used to define answer spaces and improve navigation (Jenkins and Dillman, 1997). Respondent friendly design improved response rates in national tests by about 3 percentage points (Dillman, Sinclair, and Clark, 1995). The research did not examine the effects of format changes on race and Hispanic origin data, but improvements in item response rates were expected. Additional graphics design changes (an official Census 2000 logo, icons illustrating census uses, color) were introduced in the hope of boosting response, and the form was shortened by providing space for fewer people per household than in 1990.

Fourth, *race categories were modified*. The OMB split the 1990 "Asian or Pacific Islander" category into "Asian" and "Native Hawaiian or Other Pacific Islander" in 2000. "Hawaiian" was changed to "Native Hawaiian," and "Other Asian" and "Other Pacific Islander" were offered separately rather than as a combined category. Asian categories were alphabetized. Separate categories for "Eskimo" and "Aleut" were eliminated, and "Alaska Native" was added to the American Indian category. Based on a recommendation of the Census Advisory Committee on the American Indian and Alaska Native Populations, "American Indian" was spelled out rather than abbreviated "Indian (Amer.)" as in 1990. A separate write-in space was added for the Some other race category. The effects of category changes are unknown and expected to be slight, assuming specific races can be collapsed to comparable categories in both forms.

Fifth, *question wording changes* were introduced. The race item was rephrased as a question, and the wording of the Hispanic origin item was changed from "Is this person of Spanish/Hispanic origin?" in 1990 to "Is this person Spanish/Hispanic/Latino?" in 2000. In 1990, but not 2000, the form included examples of "other Spanish/Hispanic" groups and "other Asian or Pacific Islander" groups next to the write-in spaces for these entries. The effect of the wording changes was expected to be slight. Dropping the examples may affect reporting of specific groups.

The purpose of this report is to evaluate the combined effects of these changes on race and Hispanic origin reporting, by administering the 1990 and 2000-style forms to samples of randomly selected households during Census 2000. This experiment makes it possible to attribute differences (within the limits of sampling error) in responses provided by the two samples to the effects of the questionnaire, and to rule out the effects of population changes between 1990 and 2000 and of differences in the way the censuses were conducted. The design of the experiment does not permit estimates of the separate effects of specific design features, although prior research often sheds light on which design feature accounts for data differences.

2. METHOD

2.1 Sample design

This report compares two short form mail questionnaire treatments that were administered in Census 2000 as part the Alternative Questionnaire Experiment 2000 (AQE) and the Response Mode and Incentives Experiment (RMIE).

The AQE sample of approximately 15,000 addresses received either a 1990-style short form questionnaire (about 10,000 households) or a Census 2000-style short form questionnaire (about 5,000 addresses). Sample cases were distributed equally between high coverage areas (HCAs), which are expected to have low proportions of minorities and renters, and low coverage areas (LCAs), which are expected to have a high proportion of minorities and renters. (This implies that addresses in the LCAs were sampled with a higher probability of selection than addresses in the HCAs.)

To increase sample size and improve reliability, the AQE control panel was supplemented with mail returns from the control panel for the Response Mode and Incentives Experiment (RMIE) (Guarino, 2001). These households also received Census 2000 mail short form questionnaires, just as the AQE control panel did. The RMIE control group sample of approximately 20,000 addresses was selected from the same universe using the same stratification, except the sample was allocated proportionately to the HCA and LCA strata. This implies that addresses in the two strata had equal probabilities of sample selection. All addresses in the RMIE control group received Census 2000 short form questionnaires.

Addresses on the Decennial Master Address File in the mailout/mailback areas of the country at the time sample selection took place served as the universe for sample selection (Woltman, 1999). Consequently, addresses in non-mailback areas (mostly rural areas, either where the forms are dropped off or where the housing units are listed at the time of personal visit enumeration) were excluded from sample. This excludes certain population groups of interest for this analysis, including American Indians living on reservations and Alaska Natives. Addresses that were added later as a result of coverage improvement operations were not included because they were not available at the time of sample selection. Addresses in the Accuracy and Coverage Evaluation were excluded from sample so as not to overburden these households. A systematic sample by state, stratum (the high coverage and low coverage areas), and treatment was selected.

2.2 Experimental treatments

The following treatments were compared in order to evaluate the combined effects of questionnaire changes on race and Hispanic responses:

2.2.1 Census 2000 treatment

Census 2000-style mail short form questionnaires were mailed to households designated for the Census 2000 treatment. The forms were identical to those used in Census 2000; see Figure 1.

2.2.2 1990 treatment

The 1990-style form preserves 1990 question wording, categories, order, type size, matrix format, etc. but incorporates some recognizable elements of the 2000 design (color, logo, "Start here" instruction, envelope and letter). Any questions not included in the Census 2000 short form, such as marital status, were dropped. Figures 2 and 3 (in Appendix 1) show facsimiles of the 1990-style form, and Appendix 1 summarizes the design features that differ between the two forms.

The questionnaires were mailed out according to the Census 2000 schedule, with every sampled address mailed an advance letter, a questionnaire, and a follow-up postcard. For respondents in the AQE or the RMIE, the responses provided on the mail forms were their census data. Telephone Questionnaire Assistance operators were trained to answer questions about the instruction (in the 1990-style form) to select one race category from respondents who wanted to report more than one. Households which did not return a mail questionnaire were followed up as part of the Census 2000 nonresponse operation (or, in the RMIE, using special nonresponse procedures). They are not included in this analysis.

2.3 Data coding and processing

Except for the form differences, all experimental cases were administered and processed in the same manner.

Questionnaires from both treatments were mailed back to the National Processing Office in Jeffersonville, Indiana, where they were keyed and processed. (Production Census 2000 data were returned to the geographically designated processing office, where they were imaged.) Data for both forms were edited by applying a simplified version of the pre-edits used in Census 2000 production. (Appendix 2 summarizes the coding and pre-edit procedures.) A minimum amount of information must be present to count as a valid enumeration of a person (two of six short form items, including name). Analysis is based on 57,339 valid person records: 40,723 on 2000-style forms and 16,616 valid persons on 1990-style forms. Race data were coded and pre-edited using a simplified version of Census 2000 procedures (Census Bureau, 2000; see Appendix 2). Write-in responses were coded to determine whether they represent a valid race (and if so, which race or races) or are redundant, erroneous (e.g., a person's name is occasionally written in), fictitious or uncodable (e.g., "human") answers. In general, a write-in takes precedence over a checked box when it is inconsistent with the box, but both write-ins and marked boxes are used to classify

race. Similarly, write-in responses in the Hispanic origin item were coded and used along with the check-boxes to classify Hispanic origin. Missing data were not imputed or allocated, as they would be in fully edited census data. In 1990, but not 2000, a content edit followup operation was conducted to obtain more complete responses in households which provided insufficient data.

2.4 Analysis

All cases are weighted to reflect correct sampling probabilities by stratum, and are nationally representative of areas in the mailout-mailback universe. Standard errors and t-statistics are computed using VPLX's stratified jackknife replication method (Fay, 1998) to take account of the stratified design and the clustering of people within households. The report uses $\alpha = .05$, but also indicates differences significant at the .10 level. Standard errors are given in parentheses in the tables.

3. LIMITATIONS

Results of the experiment are generalizable only to the Census 2000 mailout-mailback universe. Excluded are mail nonrespondents enumerated in nonresponse followup, and segments of the population enumerated in other operations (such as American Indians on reservations and Alaska Natives).

The design of the experiment does not permit estimation of separate effects of specific design features.

The sample size is relatively small, so statistical inferences about small differences between forms, or small population groups (such as detailed Hispanic groups) may not be reliable.

A simplified, automated version of the Census 2000 coding and pre-editing procedures was applied to data from both treatments. Different procedures were used in the 1990 census, so data from the 1990-style questionnaires were not pre-edited and coded as they would have been in 1990. Missing data were not imputed or edited.

Differences in coding, pre-editing, and processing may result in differences between results reported here and 1990 or 2000 census data. Thus, these results can support conclusions about differences between 2000-style and 1990-style mail questionnaires in the *quality and content of response data* they produce, but cannot be used to draw conclusions about differences in *final data quality*.

4. RESULTS

4.1 Mail Return Rates

The rates in Table 1 are weighted and exclude undeliverable addresses and duplicate forms (Dajani and Scaggs, 2001; Guarino, 2001). Blank forms (defined as households having less than two answers for the first two persons) were treated as nonresponses.

Of the 10,499 1990-style questionnaires mailed out, 72.6 percent (excluding undeliverable addresses) were returned, while 73.1 percent of the 5,252 households in the AQE control panel returned 2000-style questionnaires. Of the 19,639 households in the RMIE panel, 12,787 or 71.5 percent returned Census 2000 questionnaires as of April 26, 2000, when the nonresponse universe was identified (Guarino, 2001). Return rates do not differ between 1990-style and 2000-style panels for the AQE. The return rate for the RMIE panel appears slightly lower than either AQE panel, perhaps because the return rate calculations for the RMIE panel exclude mail returns after April 26th, 2000. (AQE mail returns were accepted through late May or early June.)

Weighted return rates for experimental panels, by stratum (Table 1)

Panel	N of responding households	All areas	Stratum	
			HCA	LCA
1990-style (AQE)	6,357	72.6%	76.1%	57.6%
Census 2000 (AQE)	3,253	73.1%	75.9%	60.8%
Census 2000 (RMIE)	12,787	71.5%	74.8%	58.2%

There is no difference in return rates for the AQE panels in the HCA stratum, but there is in the LCA stratum. The Census 2000 panel had a higher return rate (by 3.2 percentage points, $p<.05$) than the 1990-style panel. Within each stratum, the RMIE panel had slightly lower return rates than the Census 2000 AQE panel[1].

The slight differences among the return rates are probably due to slight differences in the calculations for the RMIE and the AQE. In general, return rates for all three panels are very close, overall and within stratum.[2] We conclude that return rates for the 1990-style and 2000-style forms differ slightly, if at all, and should not bias panel comparisons.

Census 2000 AQE and RMIE panels are combined for analysis.

---

[1] The RMIE stratum return rates are calculated using a different algorithm to identify blank forms, and hence are not exactly comparable to the other rates reported in Table 1. If the same algorithm were used, the effect would be to increase the RMIE stratum return rates very slightly.

[2] Only the differences between the two AQE panels could be tested for statistical significance.

4.2 Reporting of Hispanic Origin

Table 2 presents the distribution of Hispanic origin by form, after coding and pre-editing as described above and in Appendix 2, and including missing data. Data are missing if no box is checked, and no codable write-in entry is present.

Percentage of people reporting as Hispanic in mail questionnaires in Census 2000 AQE, by form type (Table 2)

	Form type		
	2000-style	1990-style	$t_{2000\text{-}1990}$
TOTAL	**100.00%**	**100.00%**	
Hispanic	11.17 (.2928)	11.14 (.4510)	.05
Non-Hispanic	85.50 (.3153)	74.39 (.6217)	15.8**
Missing	3.33 (.1396)	14.46 (.4891)	-21.9**

**p<.05

Table 2 shows that nearly identical fractions of people were reported as Hispanic in 2000 and 1990-style forms—11.17 and 11.14 percent respectively. The fraction reported as not Hispanic is much larger in the 2000-style questionnaire, while the rate of missing data in 2000-style forms is one quarter of the rate in 1990-style forms. In past censuses, most people for whom origin is missing have been non-Hispanic. Under this assumption, the results suggest the 2000-style questionnaire did not affect reporting as Hispanic, except to reduce the number of non-Hispanics who would have left the item blank in a 1990-style questionnaire. However, the distributional effect ultimately would depend on how the missing data were edited and imputed.

The difference in rates of missing data is very large, and was expected based on previous tests of the effects of item sequence and an added instruction. In the 1990 census, the rate of missing data would not have been as high as shown in Table 2, because a content edit followup operation would have obtained the missing information for a sample of cases (10 percent of mail return short form content edit failures went to followup).

4.3 Reporting of Detailed Hispanic Origin

After Census 2000, questions arose about whether dropping the examples that appeared in the Hispanic origin item in the 1990 census (see Figs. 1 and 3) may have resulted in less complete

identification of groups such as Salvadorans and Guatemalans in Census 2000. In 1990, examples were printed above the box for "other" write-ins:
"Yes, other Spanish/Hispanic (Print one group, for example: Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.)"
In 2000, the examples were dropped:
"Yes, other Spanish/Hispanic/Latino— *Print group*" Examples may have affected reporting because they illustrated the intended specificity of response. They may also have stimulated reporting of the specific example groups.

These possible effects are examined in Table 3, which shows form differences in Hispanics' reports of membership in detailed groups. Such reports may be given by checking off one of the three boxes associated with a specific group, or by printing a group in the space next to "other Spanish/Hispanic/Latino." In Table 3, Hispanic write-in or check-box entries are classified into four categories:

- groups with *check boxes* (Mexican, Puerto Rican, Cuban), for which specific cues appear in both forms;
- groups listed as *examples* in the 1990 but not the 2000-style form (Argentinian, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard);
- *all other specific groups* with no check boxes and not listed as examples, for which cues appear in neither form; and
- *write-ins of general descriptors*, such as "Hispanic," "Latino," or "Spanish."

In addition, some write-in entries were blank or uncodable.

Detailed Hispanic Origin, by form type (Table 3)

	2000-style	1990-style	$t_{2000\text{-}1990}$
Total persons identified as Hispanic	**100.00%**	**100.00%**	
"Check box groups": Hispanic groups with separate check boxes in both forms (sum of 1-3)	70.25% (1.25)	73.23% (1.77)	-1.37
1 Mexican, Chicano, Mexican Am.	54.26% (1.38)	58.68% (2.02)	-1.81*
2 Puerto Rican	11.42% (.83)	11.01% (1.28)	.27
3 Cuban	4.58% (.54)	3.54% (.67)	1.21
"Example groups": listed as examples in 1990-style form but not Census 2000 (sum of 4-9)	6.41% (.63)	11.16% (1.17)	-3.58*
4 Argentinian	.24% (.10)	.32% (.15)	-.45
5 Colombian	1.34% (.28)	1.89% (.42)	-1.08
6 Dominican	2.59% (.43)	2.76% (.63)	-.22
7 Nicaraguan	.52% (.17)	.57% (.19)	-.21
8 Salvadoran	1.39% (.31)	2.28% (.49)	-1.52
9 Spaniard	.32% (.12)	3.33% (.73)	-4.06*
All other specific Hispanic groups	4.20 (.50)	8.68% (1.23)	-3.38*
Write-in is general descriptor ("Hispanic" / "Latino" / "Spanish")	11.90% (.88)	1.90% (.42)	10.32*
Hispanic, no write-in (or write-in uncodable)	7.25% (.66)	5.03% (.79)	2.15*
Unweighted N	5,163	3,091	

*difference between forms significant at $p < .05$

The overall fraction of Hispanics who checked the Mexican, Puerto Rican, or Cuban box (or who wrote in one of these groups) does not differ significantly between forms (70.25 percent and 73.23 percent in the 1990 and 2000-style forms, respectively). However, significantly fewer Hispanics checked the Mexican box (or wrote in Mexican) in 2000-style forms than in the 1990-style forms. This difference is probably not due to the effects of examples or the wording of the response category, which are identical in both forms ("Yes, Mexican, Mexican-Am., Chicano"). It may be a question wording effect resulting from dropping the word "origin" in the Census 2000 questionnaire. It is possible that some people who have origins in Mexico do not self-identify as "Mexican" in the sense implied by the Census 2000 question wording.

Overall, significantly more Hispanics reported in one of the "example groups" in the 1990-style form (11.16 percent, compared to 6.41 percent in the 2000-style form). Most of the difference, however, is due to a large difference in reporting of "Spaniard" (.32% reported "Spaniard" in 2000-style forms compared to 3.33% in 1990-style forms). Excluding reports of "Spaniard," 6.08% reported an "example group" in 2000-style forms, compared to 7.82% in 1990-style forms ($t=1.56$, $p<.10$). Except for the difference in reports of "Spaniard," none of the form differences for specific example groups is statistically significant at the .05 level. More Hispanics report as Salvadoran in the 1990-style form (2.28 percent compared to 1.39 percent in the 2000-style form); the difference is significant at the .10 level in a one-tailed t-test ($t = 1.52$).

Finally, significantly larger numbers of Hispanics reported in one of the remaining non-checkbox, non-example groups in 1990-style forms (8.68 percent compared to 4.20 percent in 2000-style forms).

For three categories of Hispanic groups (those with separate check boxes, those listed as examples, and the remaining groups), then, the 1990-style form elicited more reports of specific Hispanic groups than the 2000-style questionnaire. The consistency of the effect suggests that the examples improved respondents' understanding that a specific response was intended. Overall, about 92 percent of Hispanics reported a specific group in 1990-style forms, compared with 80 percent who filled out 2000-style forms. In the latter, Hispanics tended to describe their ethnicity in general rather than specific terms. About 12 percent gave Hispanic, Latino, or Spanish as their "group," compared with about 2 percent in the 1990-style questionnaire. There were also significantly more blank or uncodable write-in entries in the 2000-style questionnaire.

4.4 Race Reporting

Table 4 reports race item nonresponse rates, by form type and Hispanic origin. The first row shows that, overall, race is missing at a lower rate in 2000-style forms than in 1990-style forms. (Race is missing if no box is checked and no codable write-in entry is present.) Race item nonresponse rates are significantly lower for both Hispanics (20.79 percent compared to 30.53 percent) and for non-Hispanics (.60 percent compared to 1.5 percent). Race nonresponse is higher in 2000-style forms for people who were also missing information on Hispanic origin. (There are many fewer such people in 2000-style forms, as shown in Table 2.)

Race nonresponse rates by form type and Hispanic origin (Table 4)

	% of people missing data on race		
Hispanic Origin	**2000-style**	**1990-style**	$t_{2000-1990}$
Total population	3.27% (.1590)	5.95% (.3265)	-7.34**
Hispanics	20.79% (1.1361)	30.53% (1.8871)	-4.42**
Non-Hispanics	.60% (.0580)	1.53% (.1756)	-5.03**
Hispanic origin missing	13.18% (1.3853)	9.72% (1.0462)	2.00**

**p<.05

More complete response to the race item in the 2000-style form is unexpected. Bates et al. (1995) found the order reversal and added instruction did not affect the race nonresponse rate.

Even with the reduction in item nonresponse compared to the 1990-style form, race nonresponse remains very high for Hispanics, who are far more likely to leave the item blank than non-Hispanics.

Table 5 presents distributions by form of the five major race groups—White, Black, American Indian and Alaska Native, Asian, and Native Hawaiian and Other Pacific Islander—and Some other race. Multiple responses are combined in a "Two or more races" category. (Multiple responses within a major category, such as Vietnamese and Chinese, are classified as single race reports.)

Missing or uncodable responses are excluded from Tables 5-7. These distributions thus approximate distributions that would be obtained were missing data imputed.

Race, by Form Type (Table 5)

	Form type		t-statistic
	2000-style	**1990-style**	$t_{2000-1990}$
TOTAL	**100.00%**	**100.00%**	
White	78.21	78.93	-1.018
	(.3719)	(.5893)	
Black	11.35	11.22	.250
	(.2847)	(.4231)	
American Indian and Alaska Native	.48	.50	-.230
	(.0549)	(.0776)	
Asian	4.04	4.06	-.033
	(.1884)	(.3282)	
Native Hawaiian and Other Pacific Islander	.17	.05	2.33**
	(.0428)	(.0246)	
Some other race	3.72	4.42	-1.97**
	(.1871	(.2992)	
Two or more races	2.03	.82	7.86**
	(.1131)	(.1045)	

**p<.05

Table 5 shows three statistically significant form effects. First, as expected, reports of two or more races are more numerous in 2000-style questionnaires, due to the new "one or more" instruction. Nearly 1 percent report two or more races in the 1990-style form, however, despite the instruction to report one. In the 1990 census, multiple reports would have been edited to a single race category.

Second, the Native Hawaiian and Other Pacific Islander category, while tiny, is larger in the 2000-style forms than in the 1990-style form. This may be artifactual. The combined "Other Asian and Pacific Islander" category in the 1990 form was split into two in the Census 2000 form. People who marked "Other API" in the 1990-style form with no write-in entry are counted in Table 5 as Asians, but some may be Pacific Islanders. It is also possible that the questionnaire design changes helped Pacific Islanders find a category to identify their race.

Third, the percentage reported as Some other race is lower in 2000-style forms, consistent with research on effects of item sequence and adding an instruction. Contrary to what might have been expected, there is little or no evidence that the "one or more" option reduced single race reporting in the five major categories. There is a very slight, statistically insignificant reduction in the percentage reported as White. The percentages identifying with the major race groups are nearly the same or higher in the 2000-style questionnaire.

Tables 6 and 7 show that negligible distributional differences at the aggregate level mask some larger effects for Hispanics and non-Hispanics.

Race, by Form Type: Hispanics (Table 6)

	Form type		t-statistic
	2000-style	1990-style	$t_{2000-1990}$
TOTAL	**100.00%**	**100.00%**	
White	48.98	39.88	3.23**
	(1.5656)	(2.3463)	
Black	2.07	2.32	-.34
	(.3719)	(.6003)	
American Indian and Alaska Native	1.48	.72	1.61
	(.3767)	(.2900)	
Asian	.58	.88	-.60
	(.2219)	(.4309)	
Native Hawaiian and Other Pacific Islander	.01	.15	-1.14
	(.0072)	(.1212)	
Some other race	39.03	51.47	-4.32**
	(1.5565)	(2.4192)	
Two or more races	7.84	4.59	2.88**
	(.7311)	(.8595)	

**p<.05

Table 6 shows that 48.98 percent of Hispanics are reported as White in 2000-style forms, compared with 39.88 percent in 1990-style forms. By the same difference of about 10 percentage points, reports of Some other race are lower, 39.03 percent versus 51.47 percent. These large differences are probably due to the effects of reversing the order of Hispanic and race items, as well as the "one or more" option. The results are consistent with earlier research showing that reversing the sequence of race and Hispanic origin increased Hispanic reporting in White race and reduced reporting in Some other race.

The 2000-style form also elicits more reports of American Indian among Hispanics, although the difference is not statistically significant at the .10 level in a two-tailed test. (The difference is statistically significant for the LCA stratum, in which 2.08 and .79 percent identified as American Indian in the 2000 and 1990-style forms, respectively; these results are not shown.) The difference may be due to South and Central American Indians more readily identifying with "American Indian" than with the less clear "Indian (Amer.)" in the 1990-style form.

Finally, Table 7 shows a different pattern of form differences for non-Hispanics and those whose origin is not ascertained. Reports of White race are slightly lower ($p<.10$) in 2000-style forms, apparently due to the option of reporting more than one race. The percentages reporting as Black, Asian, or Some other race do not differ between forms. A larger fraction report as Native Hawaiian and Other Pacific Islander in 2000-style forms. A slightly smaller fraction report as American Indian and Alaska Native in 2000-style forms, but the difference is insignificant, perhaps due to the small sample size for this group.

Race, by Form Type: Non-Hispanics or Hispanic Origin not ascertained (Table 7)

	Form type		t-statistic
	2000-style	**1990-style**	$t_{2000-1990}$
TOTAL	**100.00%**	**100.00%**	
White	81.15	82.43	-1.87*
	(.3669)	(.5682)	
Black	12.28	12.02	.47
	(.3066)	(.4539)	
American Indian and Alaska Native	.38	.48	-1.12
	(.0461)	(.0805)	
Asian	4.39	4.34	.12
	(.2052)	(.3542)	
Native Hawaiian and Other Pacific Islander	.18	.04	2.74**
	(.0471)	(.0195)	
Some other race	.17	.20	-.52
	(.0304)	(.0581)	
Two or more races	1.45	.48	7.56**
	(.0980)	(.0819)	

*p<.10 **p<.05

5. CONCLUSIONS AND RECOMMENDATIONS

Census 2000 questionnaire changes substantially improve the completeness of race and Hispanic origin reporting in mail questionnaires, compared to the 1990 design. In addition, the Census 2000 questionnaire design affects race reporting. Reports of two or more races more than double in response to the "mark one or more" instruction. There are more reports of Native Hawaiian and Other Pacific Islander race, and fewer reports of Some other race.

There is surprisingly little evidence that allowing respondents to report more than one race

reduces single race reporting in the 5 major categories. The exception is a reduction in reporting of White by non-Hispanics.

For some race groups, an absence of form differences at the aggregate level masks differential effects for Hispanics and non-Hispanics. Compared to 1990-style forms, 2000-style forms elicit more reports of White race among Hispanics (the probable effect of the reversed item sequence), and fewer among non-Hispanics (probably due to the "one or more" option), resulting in no overall form difference in the fraction reported as White. The data hint at increased reporting as American Indian and Alaska Native by Hispanics and reduced reporting by non-Hispanics in 2000-style forms, but samples are too small to be sure. There is also the suggestion of reduced reporting as Native Hawaiian and Other Pacific Islander by Hispanics and increased reporting by non-Hispanics in 2000-style forms, but only the latter difference is statistically significant.

These results imply that the questionnaire changes made in Census 2000 had different effects upon race reporting by Hispanics and non-Hispanics. These differential questionnaire effects merit additional investigation, first to determine their reliability and second, to evaluate their causes.

- **Recommendation: Conduct additional research into the reliability and causes of differential form effects on race reporting by Hispanics and non-Hispanics.**

The effects of questionnaire design changes on Hispanic race reporting are fairly dramatic. Reporting as White increases 10 percentage points, and reporting as Some other race decreases by the same amount in Census 2000-style forms. This result reflects the "one or more" option and the reversal in item sequence, and is consistent with prior research. The results confirm the vulnerability of Hispanics' race reporting to question order and context effects. They leave open the question of how vulnerable Hispanics' (or others') race reporting is to other methodological effects, for example, mode of interviewing, which have not been evaluated using experimental designs.

- **Recommendation: Conduct additional research to develop more robust race measurement methods that are less vulnerable to methodological effects, especially for Hispanics.**

- **Recommendation: Conduct experimental research to evaluate the effects of other methodological influences on race reporting, including mode of interviewing and interviewer effects.**

Despite the reversed sequence of Hispanic origin and race and question wording differences, the percentage reporting as Hispanic appears to be identical in the two forms. This result implies that changes from 1990 to 2000 in the fraction of the population identifying as Hispanic are not due to changes in design of the mail questionnaire.

On the other hand, the experiment does offer evidence that the questionnaire affected reporting of detailed Hispanic origin. Hispanics who filled out 2000-style mail questionnaires were less likely to report a specific Hispanic group and more likely to report a general descriptor (such as Hispanic, Latino, or Spanish) than those who filled out 1990-style questionnaires. Although the cause of the effect is uncertain, it is probably due to the combined effect of question wording and the elimination of examples in the Census 2000 questionnaire. The examples next to the write-in box provided cues about the type of answer intended by the question in the 1990-style form. In the Census 2000 questionnaire, the instruction to "print group" right after the "Yes, other Spanish/Hispanic/Latino" response category may have suggested to some respondents that they should print whichever of these three terms they preferred. However, the hypothesis of example effects does not account for the higher reporting of Mexicans in the 1990-style form. This difference requires a different explanation, because the specific examples (Mexican, Mexican Am., Chicano) are identical in both forms. The wording change from "Is this person of Spanish/Hispanic origin?" to "Is this person Spanish/Hispanic/Latino?" may have contributed to the reporting difference. The Census 2000 question appears directed to an overarching identification as Hispanic (or Spanish or Latino), and the absence of specific Hispanic examples would reinforce this wording effect. Because the experiment was designed to evaluate the effects of all the wording and design differences between the 1990 and 2000 mail questionnaires, it is not well suited to isolating the causes for this or other differences.

- **Recommendation: Conduct additional research on the effects of examples on race and Hispanic reporting.**

This report is exclusively focused on the effects of questionnaire design changes on race and Hispanic reporting, holding constant the effects on the data of differences in pre-editing, coding, editing, and imputation procedures used in 1990 and 2000. The effects of these potential influences on race and Hispanic data also merit investigation.

- **Recommendation: Conduct research on the effects of changes in coding, pre-editing, editing, and imputation procedures on the comparability of race and Hispanic data**.

The questionnaire design effects documented in this report may confound comparisons of 1990 and 2000 census data. The degree of confounding cannot be inferred directly from the analysis reported here, which is restricted to mail short forms and does not employ fully edited data. However, it can be inferred from the experimental evidence that the differences in the design of 1990 and 2000 mail short forms would have resulted in an increase from the 1990 to the 2000 census in Hispanics' reporting of White race, and a decline in their reporting of detailed Hispanic groups, *in the absence of true change in the racial or ethnic composition or identifications of the population.* The percentage of Hispanics who reported as White (alone) was 51.7 in 1990 and 47.9 in 2000 (U. S. Census Bureau, 2001). The questionnaire effect would have led more Hispanics to report as White in Census 2000. Therefore, we can infer that the decline in White reporting would have been even larger had the 2000-style questionnaire not increased Hispanics' reporting as White, compared to a 1990-style questionnaire. We can also infer that any measured

decline from the 1990 to 2000 census in reporting of detailed Hispanic origins is overstated; the decline would have been less if the 2000-style questionnaire had not resulted in less detailed reporting. While it might be tempting to conclude that a decline in detailed Hispanic reporting was due to Hispanics' changing self-identifications, any such change can be attributed (at least in part) to changes in the design of the mail questionnaire. These confounding effects of questionnaire design differences must be taken into account when comparing 1990 and 2000 census data.

The potentially confounding effects of the questionnaire design changes upon comparisons between 1990 and 2000 census data could not be identified and measured without a replication study based on an experimental design.

- **Recommendation: In future censuses, conduct larger replication studies embedded in the census to evaluate and calibrate the effects on the data of questionnaire design changes (or other important changes in methods).**

Future censuses should conduct replication studies to evaluate the effects of questionnaire design changes on long form as well as short form items, and should employ larger samples than were available for the Census 2000 AQE, in order to improve estimates of questionnaire effects for small groups.

Acknowledgment

Thanks to Aref Dajani and Mary Ann Scaggs for preparing AQE files and calculating response rates, to Jennifer Guarino for sharing RMIE control panel data, to Arthur Cresce and Kevin Deardorf for consulting on the race and Hispanic codes and pre-edits, to Manuel de la Puente for checking computer programs, to Maria Cantwell for wordprocessing assistance, to Keith Bennett for fact-checking, and to Florence Abramson, Claudette Bennett, Debbie Bolton, Terri Carter, Cynthia Clark, Manuel de la Puente, Jorge del Pinal, Eleanor Gerber, David Hubble, John Iceland, Ruth Ann Killion, Sally Obenski, and Henry Woltman for reviewing earlier versions of this analysis.

References

Bates, N., Martin, E. A., DeMaio, T. J., and de la Puente, M. (1995) "Questionnaire effects on measurements of race and Hispanic origin," Journal of Official Statistics 11:433-459.

Census Bureau. (1996) Findings on Questions on Race and Hispanic Origin Tested in the 1996 National Content Survey. Population Division Working Paper No. 16. Census Bureau.

Census Bureau. (1997) Results of the 1996 Race and Ethnic Targeted Test. Population Division Working Paper No. 18. Census Bureau.

Census Bureau. (2000) Census 2000 100% Imputation Specifications. Population Division.

Census Bureau. (2001) "Population by Race and Hispanic or Latino Origin for the United States: 1990 and 2000." Census 2000 PHC-T-1. Internet release.

Dajani, A. and Scaggs, M. A. (2001) "AQE 2000 Response Rate Analysis." Unpublished memorandum, Census Bureau.

Dillman, D., Sinclair, M., and Clark, J. (1995) "Effects of Questionnaire Length, Respondent-Friendly Design, and a Difficult Question on Response Rates for Occupant Addressed Census Mail Surveys," Public Opinion Quarterly 57:289-304.

Fay, R. E. (1998) VPLX Program Documentation, Vol. 1. Census Bureau.

Gerber, E. de la Puente, M., and Levin, M. (1998) Race, Identity and New Question Options: Final Report of Cognitive Research on Race and Ethnicity. Census Bureau unpublished report.

Guarino, J. (2001) Assessing the Impact of Differential Incentives and Alternative Data Collection Modes on Census Response. Census 2000 Testing and Experimentation Program.

Jenkins, C. and Dillman, D. (1997) "Towards a Theory of Self-Administered Questionnaire Design." In Survey Measurement and Process Quality, eds. L. Lyberg, et al. New York: Wiley.

Kissam, E., Herrera, E., and Nakamoto, J. M. (1993) Hispanic Response to Census Enumeration Forms and Procedures. Report prepared by Aguirre International for the Census Bureau.

Martin, E., DeMaio, T. and Campanelli, P. (1990) "Context Effects for Census Measures of Race and Hispanic Origin." Public Opinion Quarterly 54:55-566.

McKenney, N., Bennett, C, Harrison, R., and del Pinal, J. (1993) "Evaluating Racial and Ethnic Reporting in the 1990 Census." Proceedings of the American Statistical Association.

Office of Management and Budget. (1997) "Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, Part II." Federal Register vol. 62, no. 210:58782-90.

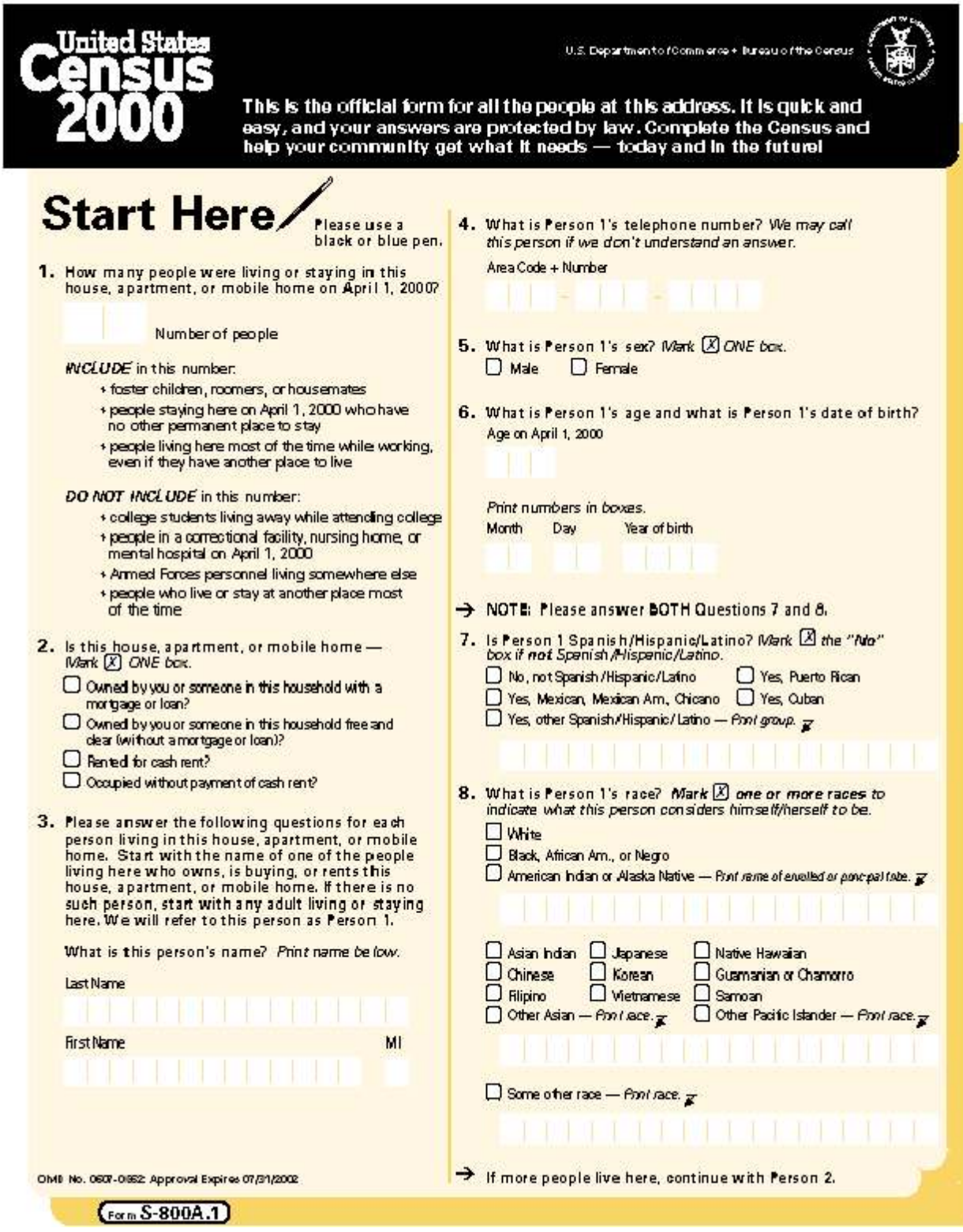
Woltman, H. 1999. Sampling Specifications for the Research and Experimentation Program in K. Shaw, Program Master Plan for the Census 2000 Alternative Questionnaire Experiment (AQE2000). U. S. Bureau of the Census, Planning Research Evaluation Division, December 22, 1999.

Appendix 1: Design Features of the 1990-Style Short Form

The intent in this experiment was to replicate features of the 1990 short form design that may affect data content and quality compared to the design of the 2000 form. The form also had to resemble, at least superficially, the Census 2000 form, which was recognizable through exposure in the advertising campaign. (See Figures 1-3 for facsimiles of the 1990-style and the Census 2000 short forms used in the experiment.)

The 1990-style form that was administered is, in effect, the heart of the 1990 short form in a 2000 shell. The experimental form preserves essential 1990 design features (question wording, order, and format) in a form which duplicates 2000 content (that is, the same questions are included) and incorporates elements of the 2000 design. The table below compares the design features of the 1990-style form with 1990 and 2000 census forms. Shading indicates which form (2000 versus 1990 census) the 1990-style form most closely resembles.

	Compared to Census 2000 Short Form	Compared to 1990 Short Form
Questionnaire Content	Identical–includes the same set of questions as Census 2000	Not comparable–marital status, whole household UHE, 7th person, many housing items eliminated
Question wording and sequence	Different–1990 wordings, categories, and sequence are used	Identical to 1990 question wordings, with minor/necessary changes: "1990" to "2000", "Sunday" to "Saturday". Due to elimination of marital status, race and Hisp. Origin are separated by one item, not two.
Question formats	Different–1990 formats are used	Matrix format is comparable to 1990, except 7th person eliminated. Question formats are identical. Format for year of birth modified slightly to allow for year 2000 births.
Instructions	Different–except the "Start Here" instruction, and the absence of an instruction book, which follow 2000.	Roughly comparable–some instructions eliminated, or minor changes. "Start Here" instruction added. Instruction book eliminated.
Structure of form	Identical–folds and size are identical to the bifold 2000 form	Different–the "flap" is eliminated; the roster is on the front page.
Color	The form uses the same colors as the 2000 form	Placement of color shading replicates 1990 use of color.
Writing implement	Different	Same–use black pencil
Other design features	Logo, heading on the front page are identical to the 2000 form. Typeface is the same as 2000.	Black registration marks and "census use only" boxes were eliminated. Type size similar to 1990.
Letter, envelope, implementation	Identical to 2000, except return envelope is yellow instead of white and is sent to J'ville	Letter is separate; in 1990, letter was the front of the q'aire. '90 envelope did not include mandatory message.



United States Census 2000

U.S. Department of Commerce • Bureau of the Census

This is the official form for all the people at this address. It is quick and easy, and your answers are protected by law. Complete the Census and help your community get what it needs — today and in the future!

Start Here — Please use a black or blue pen.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2000?

Number of people

INCLUDE in this number:

- foster children, roomers, or housemates
- people staying here on April 1, 2000 who have no other permanent place to stay
- people living here most of the time while working, even if they have another place to live

DO NOT INCLUDE in this number:

- college students living away while attending college
- people in a correctional facility, nursing home, or mental hospital on April 1, 2000
- Armed Forces personnel living somewhere else
- people who live or stay at another place most of the time

2. Is this house, apartment, or mobile home — *Mark* [X] *ONE box.*

- ☐ Owned by you or someone in this household with a mortgage or loan?
- ☐ Owned by you or someone in this household free and clear (without a mortgage or loan)?
- ☐ Rented for cash rent?
- ☐ Occupied without payment of cash rent?

3. Please answer the following questions for each person living in this house, apartment, or mobile home. Start with the name of one of the people living here who owns, is buying, or rents this house, apartment, or mobile home. If there is no such person, start with any adult living or staying here. We will refer to this person as Person 1.

What is this person's name? *Print name below.*

Last Name

First Name MI

4. What is Person 1's telephone number? *We may call this person if we don't understand an answer.*

Area Code + Number

5. What is Person 1's sex? *Mark* [X] *ONE box.*

☐ Male ☐ Female

6. What is Person 1's age and what is Person 1's date of birth?

Age on April 1, 2000

Print numbers in boxes.

Month Day Year of birth

→ NOTE: Please answer BOTH Questions 7 and 8.

7. Is Person 1 Spanish/Hispanic/Latino? *Mark* [X] *the "No" box if not Spanish/Hispanic/Latino.*

- ☐ No, not Spanish/Hispanic/Latino
- ☐ Yes, Puerto Rican
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Cuban
- ☐ Yes, other Spanish/Hispanic/Latino — *Print group.*

8. What is Person 1's race? *Mark* [X] ***one or more races*** *to indicate what this person considers himself/herself to be.*

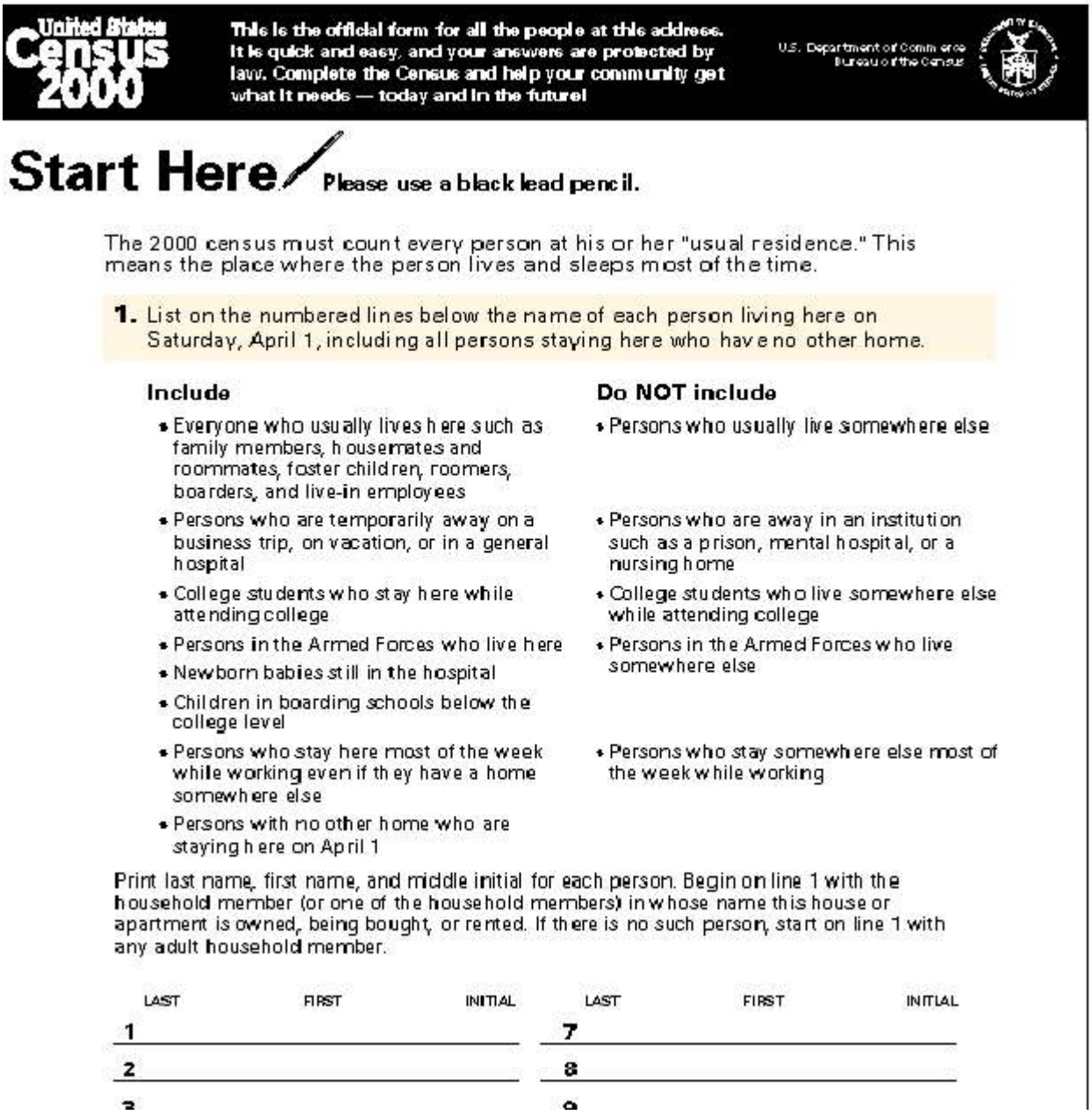
- ☐ White
- ☐ Black, African Am., or Negro
- ☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*
- ☐ Asian Indian
- ☐ Japanese
- ☐ Native Hawaiian
- ☐ Chinese
- ☐ Korean
- ☐ Guamanian or Chamorro
- ☐ Filipino
- ☐ Vietnamese
- ☐ Samoan
- ☐ Other Asian — *Print race.*
- ☐ Other Pacific Islander — *Print race.*
- ☐ Some other race — *Print race.*

→ If more people live here, continue with Person 2.

OMB No. 0607-0852 Approval Expires 07/31/2002

Form S-800A.1

Front page of Census 2000-style questionnaire (Figure 1)



United States Census 2000

This is the official form for all the people at this address. It is quick and easy, and your answers are protected by law. Complete the Census and help your community get what it needs — today and in the future!

U.S. Department of Commerce
Bureau of the Census

Start Here Please use a black lead pencil.

The 2000 census must count every person at his or her "usual residence." This means the place where the person lives and sleeps most of the time.

1. List on the numbered lines below the name of each person living here on Saturday, April 1, including all persons staying here who have no other home.

Include

- Everyone who usually lives here such as family members, housemates and roommates, foster children, roomers, boarders, and live-in employees
- Persons who are temporarily away on a business trip, on vacation, or in a general hospital
- College students who stay here while attending college
- Persons in the Armed Forces who live here
- Newborn babies still in the hospital
- Children in boarding schools below the college level
- Persons who stay here most of the week while working even if they have a home somewhere else
- Persons with no other home who are staying here on April 1

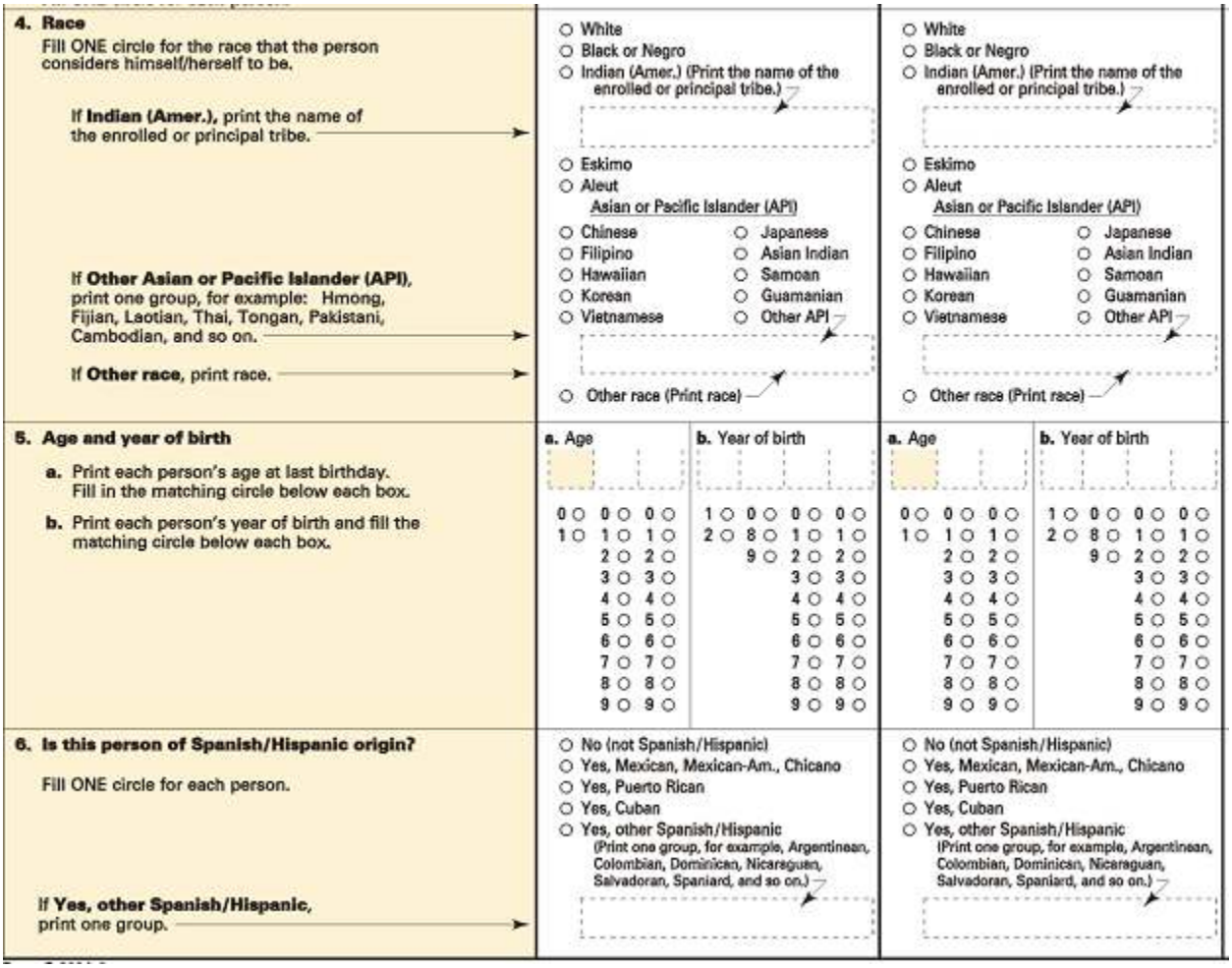
Do NOT include

- Persons who usually live somewhere else
- Persons who are away in an institution such as a prison, mental hospital, or a nursing home
- College students who live somewhere else while attending college
- Persons in the Armed Forces who live somewhere else
- Persons who stay somewhere else most of the week while working

Print last name, first name, and middle initial for each person. Begin on line 1 with the household member (or one of the household members) in whose name this house or apartment is owned, being bought, or rented. If there is no such person, start on line 1 with any adult household member.

	LAST	FIRST	INITIAL		LAST	FIRST	INITIAL
1				7			
2				8			
3				9			

Front page of 1990-style questionnaire (Figure 2)



Race and Hispanic questions in 1990-style questionnaire (Figure 3)

Appendix 2. Summary of Data Preparation, Coding and Pre-Edit Procedures Applied to Data

A. Initial file preparation. AQE and RMIE: Raw data files for the experimental panels were prepared by DSCMO using data capture specifications designed and approved for each panel. For the 1990-style panel, DSCMO developed special recoding instructions to facilitate incorporation of respondent data into production Census 2000 processing. Except for these recodes, data were entirely unedited.

RMIE control panel. Both data and programs to create the files and calculate response rates for the RMIE control panel were provided by Jennifer Guarino (PRED). The programs were modifed (i.e., to produce a person-level file and to combine all the RMIE subpanels into one control panel, rather than produce household-level files for each panel) for the different purposes of this analysis.

AQE panels: The initial AQE files were prepared by Mary Ann Scaggs and Aref Dajani (SRD), who also calculated response rates.

B. Identification of valid persons. Blank person records were not eliminated during the initial file creation, rather all 6 potential person records were retained for each form. Production census processing applies the DCAR edit to determine if sufficient data are present to represent a valid person record. A data defined person record includes at least two of the following short form items: Name (at least 3 legal characters), relationship, sex, age or date of birth, Hispanic origin, race. In the creation of the final analysis files, a simplified version of the DCAR edit was applied to eliminate blank person records and those with insufficient data. Application of the edit selected 57,339 person records for the final analysis file.

C. Correcting data capture errors.

1. For panels 2 and 4 (1990-style questionnaire), DSCMO did not capture information about multirace responses, but rather recorded a "+" when such responses appeared. There were 133 such cases. In order to capture the information, the images for the corresponding questionnaires were examined and the 133 cases were corrected to capture all write-in entries and marked boxes.

2. Inspection of questionnaire images and comparison with raw data for individual cases revealed systematic data capture errors that affected all the data for certain race categories for AQE and RMIE Census 2000 panels. In the raw data files,

RACECB07 was supposed to represent Japanese, but instead represented Other Asian.
RACECB08 was supposed to represent Korean, but instead represented Japanese.
RACECB09 was supposed to represent Vietnamese, but instead represented Korean
RACECB10 was supposed to represent Other Asian, but instead represented Vietnamese.
The data were corrected to correspond to the data capture specifications.

D. Coding and pre-editing race responses.

The raw data (corrected as described in C) contained a series of 0-1 variables corresponding to each possible race category that might have been marked. In addition, verbatim entries for all write-in spaces were captured. I pre-edited and coded these data in a fashion that somewhat simplifies but is consistent with Population Division's pre-edit and coding procedures applied to production Census 2000 race data. POP codes write-in entries into detailed race codes, which are further grouped into the 5 major race categories and Some other race (see Population Division, 2000). The criterion for allocating a specific detailed group to a major race category is the "90% rule" based on analysis of 1990 race and ancestry data. The rule is that, if 90 percent or more of a group reported as a certain race in 1990, then write-ins of that group are assigned to that race (e.g., because over 90 percent of people who reported their ancestry as Jamaican reported their race as Black in 1990, a write-in of Jamaican is classified as Black race). If a group has no dominant racial composition, it is classified as Some other race. A brief description of how the Census Bureau classifies specific groups into major race categories is as follows:

White includes write-in entries of European ethnicities (e.g., Irish, Italian) as well as Arab ethnicities (e.g., Lebanese, Syrian, Afghan).

Black includes Sub-Saharan African and Caribbean ethnicities (e.g., Ethiopian, West Indies)

American Indian and Alaska Native includes specific Indian, Alaskan, or Canadian tribes, as well as general mentions of "American Indian" or "Native American."

Asian includes Asian ethnicities or nationalities (e.g., Pakistani, Asian Indian, Japanese, Filipino).

Native Hawaiian and Pacific Islander includes Hawaiians and other groups from the Pacific Islands (e.g., Palauan, Tahitian).

Some other race includes race write-in entries of Hispanic or Latin American groups or nationalities (e.g., Chicano, Bolivian, Cuban, Spanish, Puerto Rican), groups without a dominant racial identity (e.g., mentions of United Arab Emirates, Guyanese, Moroccan, South African, Bermudan, Brazilian), and responses indicating an unspecified racial mixture (e.g., Biracial, Mulatto, Creole, Mestizo, Amerasian; but "Biracial black and white" is classified as White race and Black race, not as SOR).

Only the major race groupings were coded. The same procedures were applied to data from both the 1990-style and 2000-style questionnaires. Missing data were not imputed or edited.

The sources that were consulted during the pre-edit and coding process were the questionnaire images, accessible through FEITH software, and POP experts (in particular, Art Cresce) on the codes and pre-edit rules.

The following steps were followed:

1. Automated coding of individual write-in entries. A SAS program was written to recognize text strings in the write-in spaces, and coded them to the major race categories. This program was used to separately code multiple write-in entries for each of the three race write-in spaces (two spaces in the 1990-style form). Coding was only done to the major race groups, not to

detailed race codes. A single entry could be coded in more than one major race category (e.g. "Japanese and White" would be classified as Asian race and as White race). Variables were created to reflect the major race groups represented by all write-in entries. The development of the program was done iteratively, and the uncodable entries examined to account for misspellings and to capture and code as many meaningful responses as possible In addition, spot checking of actual responses against assigned codes was done to ensure reasonable accuracy. Questionable entries were referred to POP experts for resolution. (Certain entries, e.g., "human," "American," "pink", are considered uncodable.)

2. Generic Indians. Write-in entries of just "Indian" are ambiguous and cannot be assigned to a major race category. Such write-in entries were identified , images for the corresponding questionnaires were inspected, and codes assigned.

3. Pre-edits for consistency. A respondent's mark in a checkbox for a specific race also determines racial classification, but is usually given less priority than the write-in when the two conflict. The Census Bureau performs several pre-edits between write-in and checkbox entries for consistency. The following pre-edits were applied to these data. (Numbers in parentheses indicate the number of times the edit was performed on the combined dataset of 57,339 persons.)

a. Forms which contain generic write-ins of "Indian" were examined and classified appropriately based on information about the household as a whole. In the absence of additional information, generic "Indian" is classified as Some other race (N=15). 14 generic Indian writeins were recoded to Asian Indians, based on inspection (see 2, above).

b. In 2000-style questionnaires, if the Other Asian checkbox is marked and an entry inconsistent with Other Asian is provided in the write-in space, then the Other Asian checkbox is blanked (e.g., if Other Asian is marked but "Hispanic" is written in, Other Asian is blanked. (This is a simplification of the actual census pre-edit, which in such cases would not blank the Other Asian box if there were other persons in the household coded as an Asian race. A similar caveat applies to pre-edits 3, 4, 5, 6.) (N=39) If the Other Asian box is marked and there is no write-in entry (or the entry is uncodable, such as "human" or "American") then the Asian classification is retained.

c. A comparable pre-edit is applied to the Other Pacific Islander box and write-in for 2000-style forms (N=24).

d. In the 1990-style questionnaire, if the Other API box is marked and an entry inconsistent with Other API is provided in the write-in space, then the Other API checkbox is blanked. (N=40)

e. A pre-edit comparable to b is applied to the American Indian or Alaska Native box and write-in (N=19). (Except an entry of "Mexican" in the AI&AN write-in space would not result in the AI&AN box being blanked, but would be coded as Mexican Indian in the AI&AN category.)

f. If the Some other race box is marked and its write-in entry is inconsistent with SOR classification, the Some other race box is blanked. (E.g., if a respondent checks SOR and writes in "Polynesian", Polynesian is coded as Pacific Islander race and the SOR box is blanked.) However, if the SOR box is marked and there is no write-in entry (or the entry is uncodable, such as "human") then the SOR classification is retained. (N= 357)

g. If the Black (but not the White) box is checked and White ethnicities (e.g., English) are written in, the White ethnicities are disregarded (N=13).

h. If the White (but not the Black) box is marked and Black ethnicities (e.g., Jamaican) are

written in, the Black ethnicities are disregarded (N=1).

i. If race is blank (or uncodable) and the Hispanic origin item contains a race write-in (see below), it is used to classify race (N=14).

As noted, the pre-edits applied in this analysis are a simplified version of the actual census pre-edit and coding process. Some were not applied because there were no relevant instances in these data. (For example, in the census if all checkboxes were marked, the checkboxes would be blanked and race would be imputed.)

4. Creation of final race variables.

After coding write-ins and performing the above pre-edits, a geometric variable (RACEOMB) was created based on both the codes assigned to write-in entries and the (pre-edited) marked boxes. (Thus, for example, writing in a group classified as "American Indian" in any of the write-in spaces OR checking the American Indian box (with no write-in or an AI writein) would lead to assignment of American Indian race.) This variable captures information about all major race combinations that were reported. For this report, responses of two or more races were collapsed into a single category.

E. Coding and pre-editing Hispanic Origin responses

Hispanic origin write-in responses are also coded and used to classify detailed Hispanic group, using the Census Bureau's coding scheme. A respondent's mark in a checkbox (with certain pre-edits applied) also determines classification.

1. Coding write-in responses. All write-in responses for the item (including writeins of a major race group, if there was no Hispanic group written in) were coded into a specific Hispanic group, or (if mentioned) a major race category, using a SAS program that recognized character strings. Entries of Spanish-speaking countries or generic Hispanic or Latin entries are considered as Hispanic, while Brazilian, Portuguese, Filipino are not considered to be Hispanic. (Such entries would have been classified as Some other race reports.) If multiple groups were reported, only one was coded (which would have been the one furthest down the list of applicable character strings; preference was given to a report of a Hispanic group over a race group, and to a specific Hispanic report over a general one). (In the census, multiple Hispanic group write-ins would not have been coded in either specific category, but in a "multiple group" category.)

2. Pre-edit

a. If "other Hispanic group" is marked and a race is written in the write-in space, then the "other Hispanic" box is blanked. (N=64)

b. If "not Hispanic" is marked, but a Hispanic write-in is provided for the race item, then not Hispanic is blanked and the case is coded Hispanic. (N=63)

A pre-edit that was applied in the census but was not applied here is that if a person marked Hispanic, but had reported their race as Filipino, Brazilian, or Guamanian, Hispanic was blanked.

3. Final Hispanic origin

Based on coded write-ins and pre-edited check-boxes, a variable (HO) was created that classified respondents as Hispanic, not Hispanic, or missing. Unlike the race item, multiple responses were not allowed.